认知域 ZZ 集中体现为大数据 ZZ 和信息 DK，主要 ZZ 武器是来自三元空间的大数据，ZZ 能力体现为大数据智能分析水平。为了能够获取未来战争的领先优势，本项目拟提供战备级认知域 ZZ 地图、全球舆战态势图和全要素基础资源库、贯穿 ZZ 回路的舆战智能武器库以及平战一体的舆战试验场，为未来认知域 ZZ 提供理论保障和系统基础。

随着网络信息技术的发展，主流媒体把控舆论的能力显著下降。通过互联网、大数据、人工智能技术以及新型社会传播机理、社会心理学机制等，实现对社会群体的情绪和认知引导，已经成为国际上绝大多数国家的国家意志宣传和民众情绪管理的主要手段。在大国博弈日趋激烈的形势下，智能宣传正成为国际政治斗争的新手段、国家意识形态 DK 的主 Z 场、国际贸易 Z 的新工具，也是未来军事 ZZ 的一种新模式，并将在“信息-物理-社会”等更广泛的空间和领域发挥重要作用。

国际上，英国的剑桥分析公司除了助推 2016 年特朗普成功当选美国总统外，事实上已参与 40 多次美国政治竞选活动，并在幕后对乌克兰橙色革命、英国脱欧运动等起到了重要作用。2017 年 6 月英国牛津大学发布的一份研究报告显示，目前约有 20 多个国家在使用机器人水军影响舆论、左右时局。如图 1 所示的一系列政治事件，其背后或多或少地存在着智能技术支撑的舆论引导和民意操纵的成份。

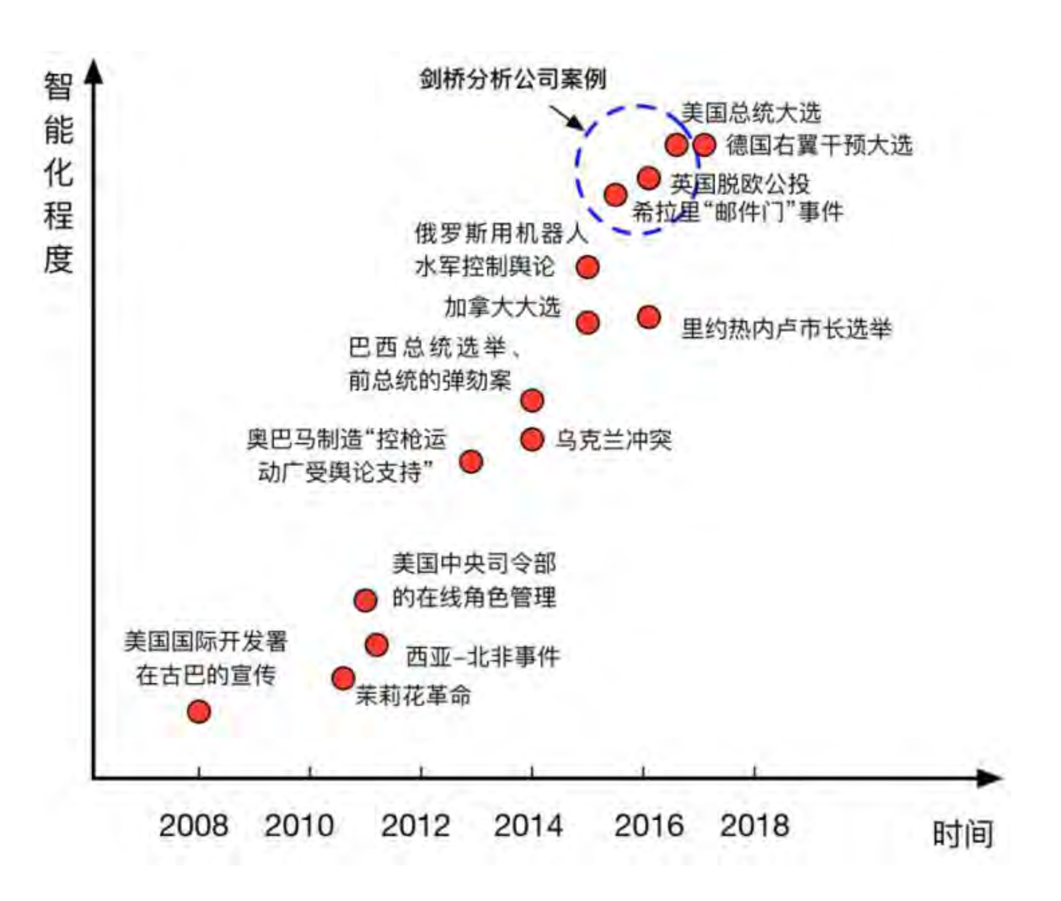


图 1 政治事件背后的的宣传智能化发展态势

西方发达国家正大力提升其舆论攻防能力。欧美等西方集团暗中支持和操控各类主打意识形态宣传、渗透、推广的网络机构与各种非政府组织（NGO），逐步发展具有一定智能性质的宣传机器人，以更有效地干涉他国政局。早在 2008 年，美国国际开发署在古巴建立了虚假社交网络，试图向潜在的持有不同政见人士宣传具有强烈政治倾向的内容，以影响古巴舆论；2010 年以来，美国利用 Facebook、Twitter 等社交网络对埃及、突尼斯、利比亚、乌克兰、委内瑞拉、叙利亚、伊朗等诸多国家的意识形态进行大规模渗透，制造了一系列颜色革命，成功导致其政局动荡，甚至政权颠覆。智能宣传正成为一种可以决定国家命运的隐形利器。鉴于其强大的舆论攻击能力，欧美等国开始制定针对性的防御计划。

美国政府在奥巴马时期，认为“网络虚假不实误导性新闻”和“俄罗斯政府通过黑客袭击与信息操控”影响到美国总统选举，将会严重威胁美国的民主政治与市场经济制度，于2016 年 12 月 23 日签署了《反宣传法案》，建立专门的反宣传中心 DK 他国的宣传渗透，同时帮助盟国获得更加真实的信息，保护其国民免受宣传洗脑，意图创造一种“能打赢思想 Z 的更全面、更积极的方式”。欧洲议会同时期也通过了《欧盟反击第三方宣传的 ZL 传播》建议案，号召欧盟各成员国采取措施反制第三方宣传。可见，欧美等西方国家在宣传领域“锐其矛”的同时，也在“固其盾”。

智能宣传将成为心理 Z 的新武器，多国正积极地筹划部署。智能化的宣传心理 Z 可以从精神层面扰乱、震慑和摧毁敌人，美国兰德公司提出了“新脑皮层 Z”的 ZZ 思想，旨在不破坏敌方组织的同时，通过信息的精准定向投放，限止和误导敌方领导者的意志力和理解力，设法影响和塑造敌方组织的行为。在该 ZZ 思想的影响下，美国 DARPA 针对新媒体的智能利用、攻击、检测和防御技术，启动了一系列面向信息心理 Z 的前瞻性研究项目，并发布了若干相应的军事 ZZ 指令（如表 1 所示)。

表 1 美国 DARPA 智能信息心理 Z 部分研究项目和条令

发布时间	名称	目标
2011.10	DARPA Narrative Networks	研究大脑如何对不同故事和言论主张做出反应，隐蔽的检测叙事对人类心理的影响及其关联神经生物学
2011.07	DARPA Social Media in Strategic Communication (SMISC)	研究利用社交媒体实现全自动或半自动的ZL沟通。
2013.01	军事信息支援 ZZ 条令 FM3-53	将意识领域的信息攻击作为一种军事活动，期望通过不易察觉的手段影响敌对政权、防止叛乱，遏制可能影响美国盟友和伙伴关系的冲突
2013.02	网络空间联合 ZZ 条令 JP3-12	提出可以对网络行为体，即网络空间的使用者“人”实施网络空间 ZZ
2015.02	DARPA The DARPA TWITTER BOT CHALLENGE	识别和清除在社交网络上进行非法信息传播来影响民众观点的社交机器人
2017.09	DARPA Active Social Engineering Defense (ASED)	自动确认、干扰和抵制通过机器人通信进行的钓鱼式攻击和社会工程学攻击

同样，俄罗斯国防部和英国国防部也提出在 Facebook 和 Twitter 等社交媒体上，通过“动态叙事”开展政治和军事宣传，对敌人进行“非致死性心理 Z”。研究显示，截至 2017 年已经有 28 个国家调用大量资源和人力来管理和操纵在线舆论，组建网络部队（如图 2 所示）。虽然各国网络部队的预算、行为和规模各不相同，但各国均已认识到，智能宣传是一种新型 Z 法，将大大拓展 ZZ 渠道、革新 ZZ 方式、提升 ZZ 效能，熟练掌握和运用智能化的舆论宣传、威慑宣传、诡诈宣传以及情感宣传武器，可达到不 Z 而屈人之兵的最高境界。

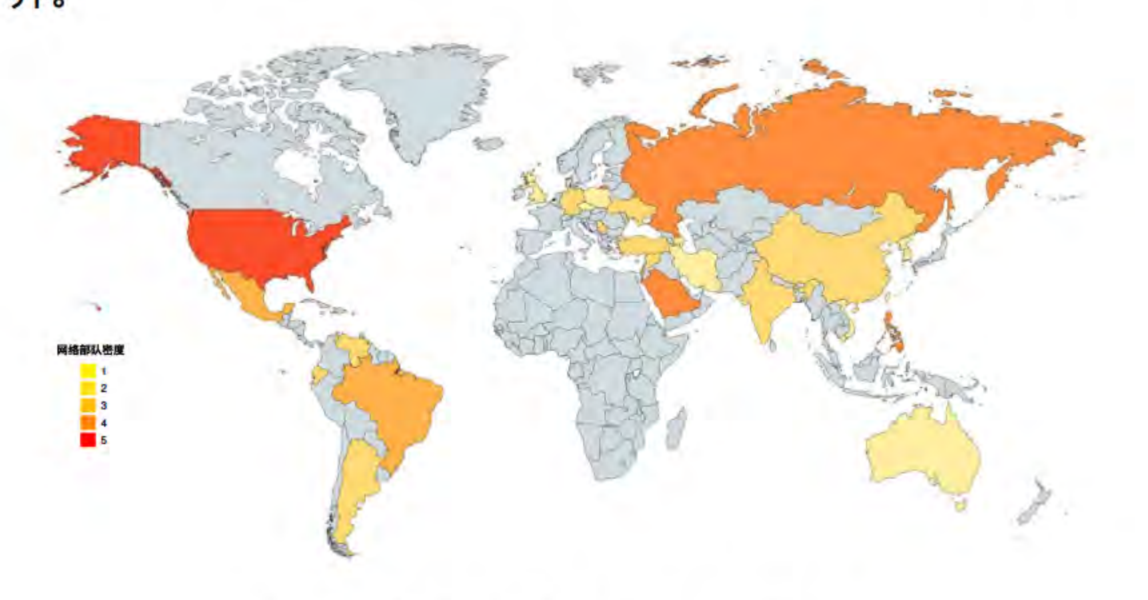


图 2 全球有组织的网络部队密度图

智能宣传将成为反华国家和势力对我舆论攻击的新工具，使国家安全面临新挑 Z。在国际舆论空间，反华势力已于各类社交媒体中组织起了规模庞大的机器人水军，包括 1989 群组、泛亚群组、独立反华机器人、反商业内容机器人、专业新闻和假新闻机器人等，主要针对历史、民族、宗

教、人权问题和舆情热点事件等，发布倾向性言论，营造舆论假象和虚假共识，混淆国际视听，抹黑我国的国际形象；高调宣扬“疆独”、“藏独”、“台独”和“港独”等分裂思想，为其分裂中国的行径制造舆情氛围。在国内舆论空间，以美国为首的西方集团在我方社交媒体平台上兴建机器人水军，窃取用户隐私、传播虚假消息，在政治、经济、文化、社会、生态等方面制造思想混乱、开展颠覆破坏活动，试图引发我国社会动荡变革。智能化的宣传工具将全方位、多领域、深层次地威胁我国国家安全。

智能宣传的强大舆论引导和动员能力，源自认知科学、大数据与人工智能技术的新发展。在认知科学层面，人们越来越注重信息、生理、心理、行为、环境的互动机制研究。相关研究表明，无意识的语义启动可以影响大脑活动，对比实验发现阈下信息刺激可使得政治观点极端的两类人政见趋于统一；网民情感能被信息改变，并通过网络进行大规模传播。近年来，人们更加注重从计算神经科学角度研究非程序的、适应性的、大脑风格的信息处理的本质和能力，这将成为智能宣传研究的重要科学基础。在大数据和人工智能技术层面，注重信息受众的海量背景数据和行为数据的收集与分析，通过数据挖掘技术支撑的心理画像模型与人工智能技术支撑的广告定向和行为分析算法，利用全方位的媒体渠道，构建规模庞大的实时宣传生态系统，对受众进行精准的信息分发和扩散，从而改变民众的心理和行为。今后，大数据、人工智能技术与计算神经科学将会深度融合，进一步提升智

能宣传的实 Z 效能。

本项目直面意识形态和国家安全领域尖锐复杂的斗争特别是“颜色革命”的现实危险，顺应网络化和智能化时代高科技发展和学科交叉融合趋势，构建智能科技支撑下的新型宣传技术体系。借助于智能宣传理论创新，创新宣传受众、内容、情境与 Z 法等要素智能分析的新型传播理论；突破一批具有精确、实时、可控和隐蔽等智能特征的共性关键技术，实现宣传工作的内容创新、手段创新和 Z 法创新；发展高度自主、实 Z 可用的智能宣传武器，针对敌对目标、亚稳定人群和普通大众的心理状态、思维意识和行为模式进行干预，并在若干重要或敏感场景、事件中开展示范应用，解决国家和军队在宣传领域的紧迫现实需求；运用人工智能在 ZL 欺骗与反欺骗中对对手精巧“设局”和巧妙“抽身”，并有机融合到联合 ZZ 体系中。

舆论是公众对于特定话题或事件的总体观点、态度和信念。舆论自从人类社会诞生之日起便在人类社会发展中扮演着重要的作用，各国政府、企业和社会团体都非常重视舆论和舆论引导。舆论引导或舆论导向，是一种运用舆论去引导人们的观点、态度和信念的信息传播和宣传行为。上升到国家政治、经济和军事层面，舆论引导则演变为更具 DK 性和体现国家博弈的舆论 Z。舆论 Z 是 Z 争双方依据传播学原理，利用电视、广播、网络、报刊等大众媒体，有计划、有目的地向受众传递经过选择的信息，宣扬己方对特定事件的立场、观点和看法，阻断、瓦解和反击敌方的舆论攻势，从而影响

受众的情感和行为，引导社会舆论、影响民意归属，造成有利于己的舆论态势。舆论 Z 是一场没有硝烟的 Z 争，被称为“不 Z 而屈人之兵”的攻心 Z，是借助舆论从内部瓦解和分化敌方的 Z 争手段，通常具有 Z 线长、高隐蔽性、强破坏性、Z 争影响难以消除的特点。

国家间的舆论 Z 是国家意志在舆论上的自然延伸和集中体现，是营造民族认同和国家博弈的重要载体之一。自古以来，舆论 Z 就是重要的 Z 争手段之一，例如楚汉之争时的“四面楚歌”之计和二 Z 时美国对德国实施的舆论 Z。冷 Z 以来，以美国为首的西方国家开始将舆论 Z 作为意识形态输出和在全球范围内实现其国家意志的重要武器，舆论 Z 在苏联解体、东欧巨变、科索沃事件、中东“颜色革命”、克里米亚事件、英国脱欧事件、美国总统大选等影响世界格局的重大政治事件中都发挥着重要的甚至不可替代的 ZL 作用。近年来，美国依托其强大的经济实力，凭借其在信息领域的技术和平台优势，以 Google、Facebook、Twitter、YouTube 等互联网平台为武器，不断在全球范围内掀起舆论 Z，对不符合美国国家利益的国家和政府实施舆论引导，作为实现其国家意志和保障其国家利益的桥头堡和先头兵。舆论 Z 的威力和重要性逐渐被世界各国所重视，开始作为保障国家利益的重要 ZLZ 场。近年来，俄罗斯、英国、法国等安理会常任理事国在舆论 Z 方面开始展示实力，舆论 Z 美国一家独大的格局开始弱化，美国也从舆论 Z 的全面出击逐渐转向攻防兼顾。我国作为世界第二大经济体和最大的发展中国家，一直

是美国舆论 Z 的主要目标之一，例如美国等国家炮制的“中国威胁论”、打着“普适价值观”的旗号进行的文化侵略、借助互联网平台输出的“历史虚无主义”和“佛系丧文化”。为此，我国亟需提升自身的舆论 Z 能力，服务于我国舆论 Z 从积极防御逐渐转向 ZL 进攻。

舆论 Z 的主要手段是舆论宣传。广义而言，舆论宣传是运用各种媒介进行观点传播，以影响人们态度、引导人们行为的一种社会性传播活动。舆论宣传本质上是一种信息传播行为，其目的是使受众知悉、认同、传播某种特定的观点。“知悉、认同、传播”从宣传目的的角度大体界定了宣传的三个阶段。第一阶段，宣传主要借助于各种媒介让大众知悉某种信息或观点，主要致力于汇聚大众的注意力和获得大众关注，更多地属于“广而告之”的范畴，主要手段和平台是大众媒体，典型的案例是新闻报道——选择具有重要性、时新性、接近性、显著性和趣味性的信息进行宣传从而实现获得关注的宣传目的。第二阶段，宣传则致力于说服受众认同和接受宣传的观点，这一阶段的宣传利用受众的社会关系、心理、文化、教育、地域、信仰等因素实现宣传目的，典型的案例是舆论引导。第三个阶段，宣传致力于在受众知悉和认同特定观点的基础上使其自发地传播该观点从而形成社会舆论，这一阶段的舆论宣传属于社会化的智能宣传，是舆论 Z 的最高境界。

舆论 Z 关系国家政治经济安全和社会稳定，已经从国家之间竞争的软力量发展成为影响国家 ZL 和发展环境的硬力

量。习近平总书记在党的新闻舆论工作座谈会上的讲话上指出的“团结稳定鼓劲、正面宣传为主”、“建设有国际影响力的外宣媒体旗舰”和最近网信工作座谈会上的讲话指出的“发挥网络引导舆论、反映民意的作用”、“网上网下要形成同心圆”对于强化正面引导、提升新闻舆论工作的吸引力、凝聚力、传播力和向心力指明了重点。实现习近平总书记提出的新闻舆论工作目标的重要途径是建设智能宣传系统，为维护国家繁荣稳定的良好舆论氛围营造和体现国家意志的对外正面宣传提供有力保障和坚强支撑。总体而言，我国目前的新闻舆论工作还主要集中在新闻宣传和舆论引导两个阶段，更多地体现在服务于保障国家经济发展和维护社会稳定方面，其技术水平和即Z力还不能完全甚至完全不能支撑大规模的舆论Z，亟需装备符合我国新时代发展需求的、支撑舆论Z的智能宣传系统技术。

以计算机和互联网为代表的信息技术革命以前所未有的深度、广度和尺度革新了舆论Z的范畴和影响力，呈现出跨空间、跨媒体、跨网络的特点。(1）跨空间：计算机和互联网极大地推动了人类社会的信息化水平，实现了人类社会、物理世界、网络空间的三元空间深度融合，大大缩短了人和人之间、信息和信息之间、人和信息之间的距离，为宣传和舆论提供了便捷的通道和广阔的空间。(2）跨媒体：文字、声音、图片、视频、表情等多模态数据提供了跨媒体的宣传媒介，以丰富的手段大大提升了信息时代宣传的冲击力。(3)跨领域：舆论不再局限于单个事件或单一领域，而是在政治、

经济、军事、文化等多个领域综合发酵形成“舆论合力”或“舆论场”，其影响也更加深远和难以消除。上述特点使得信息时代的舆论Z在时间、空间、传播媒介、影响范围等多个维度上呈现出不同于传统舆论Z的挑Z。

伴随着信息革命的飞速发展，互联网、万维网、社交网络、社交媒体、移动互联网及各类“互联网+”服务应用，正在全面改变人们的生产生活方式，深刻影响着人类社会历史发展进程。网络空间是信息传播的新渠道、社会治理的新平台、国家主权的新疆域和舆论宣传的主Z场。借助网络空间的舆论宣传和渗透危害政治安全、威胁经济安全、侵蚀文化安全、破坏社会安定，网络空间的舆论Z方兴未艾，一场争夺网络空间舆论主动权的竞争日趋激烈，现在舆论Z的跨空间、跨媒体、跨空间的特点使得传统的舆论宣传手段不再适用，亟需新的基于人工智能和大数据的智能宣传系统技术。

智能宣传系统技术需要从三个层面形成突破：（1）应用范畴层面,需要从国内的非DK环境到走向全球的DK环境。传统的舆情系统更多地侧重于舆情监测，服务于国家公共安全和社会稳定，智能宣传系统则更多地侧重于舆论引导，服务于国家意志输出和国家博弈的制信息权；（2）技术手段层面，需要态势感知走向精准导控。传统的舆情系统更多地是采用信息技术实现对舆情信息的自动获取和分析处理，属于态势感知层面，而智能宣传系统则致力于采用大数据与人工智能技术实现集态势研判、目标发现、突破点选择、高聚能精准打击和效果评估为一体的“闭环”系统；（3）呈现形式

层面，从软件系统走向复杂智能体。无论是数据感知获取、信息传播预测、群体画像和身份标识等，传统的舆情系统在呈现形式上都集中在技术和功能层面。智能宣传系统则是一个体系化的、浸入式的舆情引导系统，以一个复杂智能体的形式存在。

智能宣传系统技术的研究和部署是一个复杂的社会工程，涉及多个学科，包括计算机科学、人工智能、心理学、计算社会学、传播学、系统科学、博弈论和社会工程学。计算机科学将从计算层面解决数据的感知获取、存储管理和分析处理；人工智能解决系统的智能化实现，包括机器学习技术、深度学习技术和强化学习技术；心理学则关注受众在接受舆论宣传时的心理活动规律，包括从众心理、一致性心理、逆反心理、标新立异等；计算社会学则重点从社会层面探索文化、语言、地域、种族等社会因素对舆论宣传的影响机制；传播学探索舆论宣传的宏观传播态势和微观传播机理，支持传播模式的设计和传播接入的时机；系统科学则从复杂系统的角度探索通过多智能体平行社会对效果进行沙盘推演和效能评估；博弈论从机制设计的角度服务于 DK 环境下舆论 Z 的实施；社会工程学从心理脆弱性的角度探索基于社会工程学的舆论传播与引导。

智能宣传系统技术可作为一种 ZL 武器，单独作为 Z 争的发起点和终结点，和服务于情报 Z 的舆情系统和情报系统具有本质区别。智能宣传系统是一个基于跨空间跨媒体跨领域大数据分析技术的人在环路、虚实结合、自适应的智能化

体系化 ZZ 系统。智能宣传系统包括 ZL 态势感知、目标突破口发现、高聚能精准导控和效果评估四个环节。（1）态势感知层面，研究一体化的智能宣传技术态势感知体系，包括综合线索感知、ZL 意图研判、舆论传播溯源、综合态势表达与可视化等；（2）目标发现层面，采用智能宣传技术进行目标群体发现、用户画像与行为建模、攻击路径识别、心理脆弱性分析等；（3）精准导控层面，构建具有前瞻性和颠覆性的宣传引导机制，研究信息智能生成、舆论传播与 ZL 诱导、舆论场营造、舆论传播与极化、舆论反制与渗透等；（4）效果评估层面，研究舆论决策体系构建、复杂开放环境博弈、镜像 Z 场环境演练、舆论净效能评估等。智能宣传系统技术着眼于为“天朗气清、生态良好”的舆论环境营造提供能力支撑，更为舆论 Z 提供 ZL 支持，为国家实施舆论 Z 提供 ZL 武器和 Z 术手段。

综上所述，面向舆论 Z 的智能宣传系统技术的研制和部署势在必行、正当其时，符合我国新时代实现大国和平崛起的时代需求，是提升我国综合国力和 Z 备能力的重要手段，是我国谋求新时代更大发展空间和信息 ZL 纵深的必由之路。

目标：

针对个性化推荐系统 GJ 手段隐蔽性低、GJ 行为易被推荐系统检测、GJ 策略未充分考虑推荐系统工作机制、GJ 效果单一不可持续和社交推荐场景 GJ 手段匮乏等难题，突破用户信息消费模式分析、推荐系统智能感知、推荐系统脆弱性分析、推荐系统隐蔽 CT 和社交推荐隐蔽传播通道构建等关键技术，达到用户行为模式识别效果准、推荐系统工作机制认知全、推荐系统脆弱性分析可靠性高、推荐系统 CT 通道隐蔽性强和社交传播通道覆盖范围广等指标，为形成用户信息消费行为模式识别、推荐系统工作机制白盒化、推荐系统脆弱性评估、个性化 CT 内容生成与送达以及社交机器人网络构建与传播调度等能力。

背景：

推荐系统的出现，便利了用户信息获取的过程，也为内容生产商新增了传播渠道。进一步地，内容生产商可以尝试对推荐系统进行 GJ，实现内容的定向推送。基于个性化推荐系统构建隐蔽传播通道，需要对现有的推荐系统进行 GJ，使其做出于我方有利的推荐策略。

现状：

由于推荐系统自身具有的用户参与性和信息开放性，导致系统很容易受到外界的 GJ，出于商业竞争等目的，一些恶意用户人为地将大量伪造的用户概貌（User Profiles）植入系统中，使得他们成为大量用户的最近邻，从而使用户对目标物品的偏好发生偏移，导致系统产生对他们有利的推荐。这样的 GJ 方式通常称之为用户概貌植入 GJ，因为 GJ 用户特别像我们生活中所说的“托”，因此又叫做托 GJ，这些虚假的用户概貌称之为 GJ 用户概貌（Attack Profiles）。当推荐系统受到外界 GJ 用户的植入后，其推荐结果会发生很大的变化。常见的 GJ 类型有两种：推 GJ（Push Attack）和核 GJ（Nuke Attack），前者会导致推荐系统以更高地频率推荐目标物品，而后者则是为了降低目标物品的推荐频率。

相关国外项目：

近年来美军投入数百亿美元，通过国防部、DARPA 等机构，部署了一系列从舆论数据体系、关键技术支撑平台到舆论传播和操纵的实战项目。据“棱镜”计划披露，美军通过直接进入微软、谷歌、苹果等互联网公司服务器，实现对全球媒体和舆论数据的监控；部署 XDATA、Insight 等大数据快速处理和深度分析技术研究，支撑大规模舆论作战应用；通过“社交媒体战略传播”、“网络魔术师”及“诚挚之声行

动”等舆论操纵项目，对竞争对手进行否定、破坏、贬低和欺骗，帮助美军实施对抗 GJ 和舆论宣传战。同时美军正在加强包括“深度伪造”、“真相终结”、“社交媒体机器人”等下一代舆论监测和操纵技术研究，部分成果已形成实战能力。在这些项目部署中，**操纵内容的定向送达**都是一个不可或缺的部分。

牛津大学网络研究院发起了“计算宣传项目”（The Computational Propaganda Project）。自 2012 年以来，该团队一直在调查算法、自动化和计算宣传在公共生活中的应用。政治机器人在主要社交网络应用上操纵公众舆论。这个项目使一个新的社会和信息科学家团队能够调查自动脚本（通常称为 bot）对社交媒体的影响。该团队研究了机器人脚本和制造这种机器人的人，然后与计算机科学家合作，改进捕捉和阻止这种机器人的方法。经验表明，政治机器人最有可能出现在国际危机期间，通常是为了促进陷入困境的政府的利益。政治人物利用机器人操控对话，遣散反对派，并在 Twitter、Facebook 和新浪微博等热门网站上制造虚假支持。该项目会与受委托制造机器人的政治顾问和计算机专家进行国际实地调查，建立一个涉及机器人的政治事件的原始数据库，并利用这些知识来制作更好的工具，以便在政治机器人出现时进行检测。该项目定位于实时社会和信息科学，会定期向记者、工业界和外交政策专家传播研究成果。该项目希望通过开发一个政治机器人检测和编译原始数据集的

专家网络，帮助研究人员更好地了解机器人如何操纵社会网络，同时促进社会科学、计算机科学和工业界关于问题规模和可能解决方案的讨论。

该项目组于 2017 年发表了第 12 号研究报告:《水军、喷子和麻烦制造者: 全球有组织的社交媒体操纵盘点》。该报告比较了 28 个国家的网络宣传活动，清点了网络宣传的消息种类、价值（valence）和通信策略，并根据组织形式对其项目预算和人员配置方面的能力进行了评估。报告显示，最早关于"有组织的社交媒体操纵"的报告出现在 2010 年,到 2017 年已经有 28 个国家的情况有迹可循。隶属于政府、军队或政党团队的网络部队致力于操纵社交媒体舆论，作为一个普遍的全球性现象，许多国家调用大量资源和人力来管理和操纵在线舆论，有时针对国内受众，有时针对外国公众。几乎每个威权国家都会通过社交媒体有组织地针对本国受众进行宣传，但针对外国公众的社交媒体宣传却不常见。然而并非只有威权国家才这样做，最早关于政府参与舆论推动的报道就涉及民主国家，并且政治传播技术的创新都来自政党。样本中几乎所有的民主国家都曾组织过针对外国公众的社交媒体宣传战，而针对本国选民的宣传都是由政党支持的。在社交媒体操纵背后的网络部队方面，报告指出：网络部队通常由不同的成员构成，它们使用各种策略、工具和技术进行社交媒体操纵。随着时间的推移，网络部队的组织模式已

经从涉及社交媒体舆论操纵实验的军事单位覆盖到与政府签订合同开展社交媒体宣传的战略传播公司。而且，网络部队在将来无疑将持续存在，不断演化，成为全球现象。

研究内容：

1.1. 推荐系统智能感知

现有的推荐系统 GJ 方法基本上都是围绕推荐系统的协同过滤策略展开的。实际上，推荐系统是一个复杂系统，从内容选择，到用户定位，再到信息与用户的匹配，多项机制共同作用完成最终的推荐过程。本关键技术定位于推荐系统工作机制白盒化，旨在揭示推荐系统中关键环节的运行机制，为 GJ 方法的设计提供运行机制基础。

1.2. 推荐系统脆弱性分析

当前推荐系统最常用的共同访问推荐技术、协同过滤技术以及最新的深度神经网络和图神经网络技术等，都存在着对抗 GJ 下的脆弱性问题。充分研究这些方法的脆弱性，一方面可以在信息作战中，针对对方推荐系统的脆弱性进行 GJ，实现敌为我用；另一方面，可以提升我方网络空间系统的鲁棒性和信息安全。本关键技术定位于探寻现有推荐系统的脆弱性，为 GJ 手段的设计提供 GJ 对象。

1.3. 推荐系统隐蔽 CT

在推荐系统智能感知和推荐系统脆弱性分析的基础上，进一步研究推荐系统隐蔽 CT 通道的构建，实现消息的隐蔽送达，且不被推荐系统和用户察觉。具体来说，推荐系统主要包括物料选择、用户定位和推荐机制三个方面。本关键技术分别针对这三个关键环节制定 CT 技术，最终实现对推荐系统的隐蔽 CT，完成消息的定向送达。

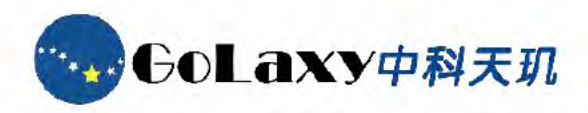
1.4. 社交推荐隐蔽传播通道构建

社交网络的角色已逐渐由网络社交向社交媒体转变。在社交平台上，用户可以选择感兴趣的用户进行关注，并后续从关注用户处获取信息。社交网络的存在，大大加速了信息的传播和送达的进程。本关键技术定位于社交推荐系统中隐蔽传播通道的构建，通过培养和模拟社交账号，实现对社交网络的覆盖和信息的定向送达。本关键技术包含如下四个子课题。



天玑数据采集中心介绍





CONTENTS

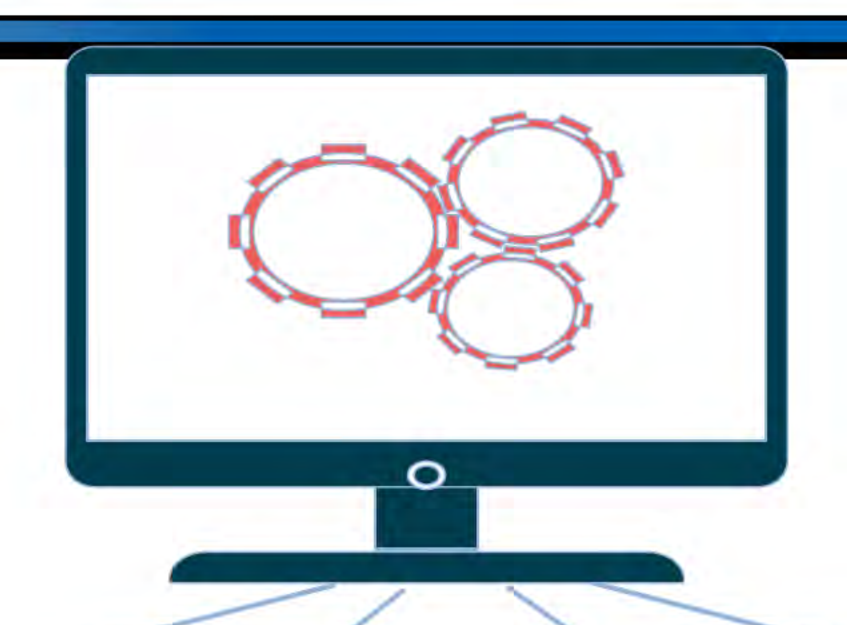




中科天玑数据采集中心致力于建设整套的数据采集方案，基于分布式高速采集技术，可一站式解决数据定制、采集、分析、挖掘、推送的问题，满足客户对即时数据、采集范围之外的历史数据及定制数据的采集需要，信源全面覆盖全网各类国内外资讯媒体、自媒体、社交网络信息，每天从千万级的信息中采集、聚合分析网络信息，实现对领域热点、重大事件、敏感事件以及用户关注事件数据自动发现，广泛服务于公检法体系、国家部委、网信、国安、军工、企事业单位等，满足客户需要。







自主研发技术

是以完全自主研发的采集技术引擎为核心，快速获取多通道规范化网络舆情数据

可拓展架构

设计可扩展的系统架构，帮助用户实现实时监测及追踪，自动采集数据，并对采集结果进行规范化处理与分析

开放接口

建立多通道统一数据系统，定义通用的数据结构以及开放的数据交换接口，实现不同舆情信息通道的融合

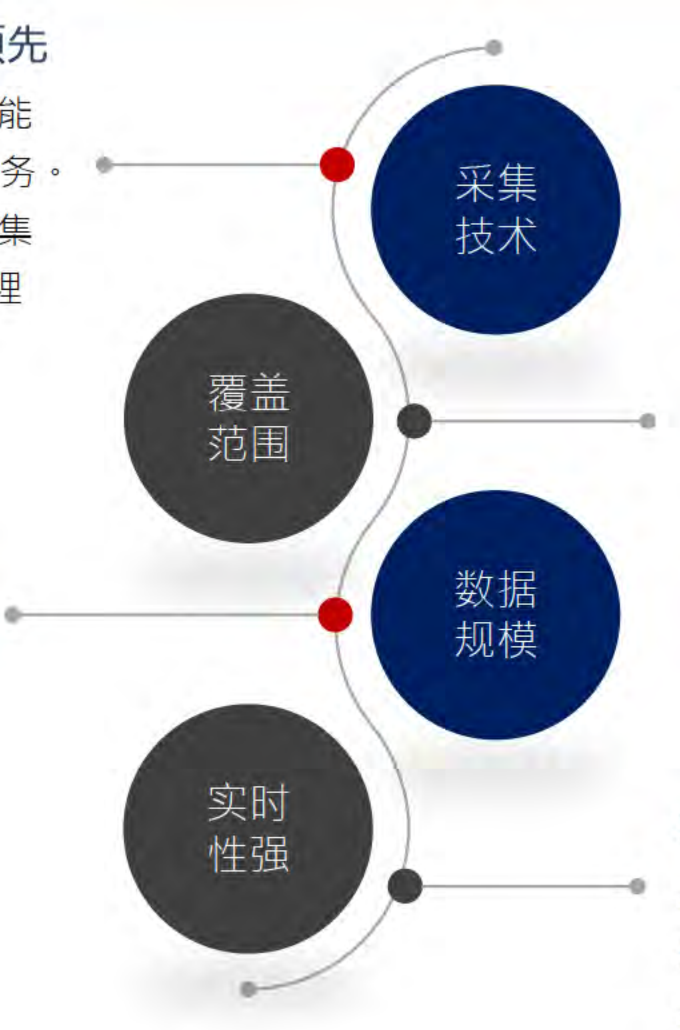
一站式平台

实现数据采集、处理、推送、分析一体化，满足客户对即时数据、采集范围之外的历史数据及定制数据的需求



采集中心简介





采集技术领先

分布式云采集架构，模块化设计，按需动态扩展，功能强大、高效稳定，百级采集服务器，7*24小时运行任务。系统支持多种采集模式，可灵活配置，多线程并发采集提高采集效率，支持JS渲染、验证码解析、动态IP代理等，有强大的反爬能力。

采集覆盖面广

系统采集信源覆盖面广，覆盖了新闻、APP新闻、境外新闻、境外社交、微信、微博、论坛、贴吧、博客、电子报、问答等10多种媒体类型，以及百度搜索、360搜索、必应搜索、搜狗搜索等主流搜索引擎。其中新闻监测的版块数达到十万级，境外新闻近万，覆盖了所有主流新闻媒体和财经媒体；论坛监测网站数达到万级；微信监测公众号规模达到百万级；微博采集为6000万+；搜索引擎覆盖了百度搜索、360搜索、必应搜索、搜狗搜索等主流搜索引擎，金融、财经、经济等多种类型，万级以上的搜索关键词。

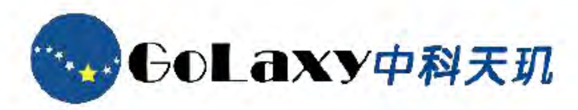
采集数据量大

系统已积累了近4年的采集数据，系统每日新增数据量6000万-8000万，其中微博为千万级，新闻、微信、论坛等其它通道百万级。

采集实时性强

系统支持分布式多节点高并发采集，支持搭配多种分层、分级策略抓取，更新时效性强，对于不同类别的数据，系统采集延迟在1-30分钟以内，并支持用户需要按需定制。

技术架构



数据

- 新闻
- 微博
- 微信
- 论坛
- 贴吧
- 博客
- Twitter
- 元搜索
- 定制采集
- ……

互联网

采集

信源管理

Zookeeper

任务管理

任务调度

采集执行器

容器

运行

服务器A

……

采集执行器

容器

运行

服务器B

容器仓库

非侵入式控制策略

IP控制

JS解析

验证码解析

频率控制

私有云(K8S)

分析处理

- 数据去重
- 数据去噪
- 数据清洗
- 信息提取
- 关键词提取
- 实体提取
- 情感分析
- 热点分析

数据加工总线

存储

- Mysql
- Redis
- Kafka
- Elasticsearch
- HDFS
- 文件

分布式存储平台

- 云采集架构
- 动态IP代理防封堵
- 验证码自动识别
- JS动态解析
- 文本语义分析
- 日志分析统计报警



核心技术



天玑数据采集中心

天玑数据采集中心

实时采集**海量**互联网信息和移动APP数据，抽取关键信息，实现多通道跨平台的大数据内容汇聚。同时提供对外数据访问接口，方便其他系统进行集成。

核心技术

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发







天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发





核心技术



统一信源管理——核心设计理念

清晰的信源层级划分,统一管理信源

- 对网络信息来源进行三级划分：信源、信簇、传媒

信源

- 信息来源，信源是指最小粒度的信源，例如新闻频道、论坛版块、微博博主等

信簇

- 信簇为信源集合。一组具有某种相似特性的信源构成一个信簇。例如一个新闻网站或者一群微博博主。

传媒

- 传媒全称为传播媒介。传媒是信源的一种属性，表示该信源传递信息的手段，例如新闻、论坛、博客、微博、微信、APP等与传播技术有关的媒体。

自动配置,快速生成高质量模板

- 提供统一的采集、抽取模板配置平台，便于自动生成配置，人工审核、修正配置

快速配置

- 通过信源自动拆分、模板自动配置模块，结合可视化标注工具，实现快速的模板配置

协同抽取

- 多种抽取算法协同合作，结合组合抽取策略，实现高质量的结构化数据提取

多阶段验证,夯实采集抽取验证结果

- 信源验证按照配置活动的各个阶段，分为入口验证，正则验证，抽取验证，在线模拟验证

入口验证 → 正则验证 → 抽取验证 → 在线模拟验证

结合业务系统数据，反馈信源状态

- 提供反馈机制，汇总线上业务系统数据
- 结合主动检测和被动检测，提升信源失效检测的时效性

信源信息统计功能的接口

- 线上的系统可以基于此接口推送信源采集量、信簇采集量、信源已采集的文章最新发布时间、采集时间

失效检测反馈的接口

- 线上的系统可以基于此接口推送潜在失效信源及失效原因



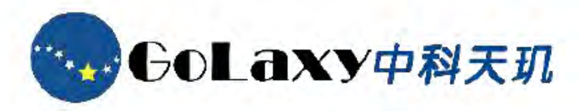


天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发



核心技术



统一信源管理——多租户

整体信源112.8万
新闻：24.5万
论坛：88.2万
电子报：1000

统一信源管理平台
本部信源管理员
烟台信源管理员
标注实习生团队
信源
烟台采集中心
ZX采集中心
WX采集中心
。。。。
数据
烟台YQ
通管YQ
南部ZQ
ZX YQ
ZX 金融
WX YQ
XW TK

目前共支撑
65
个系统

技术维护：3人
信源管理：2人
标注团队：30+





天玑数据中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发





统一信源管理——数据安全

基于Browser的配置前台

基于Java的Web Server

用户访问权限管理
强制登录限制
用户角色权限
用户口令加密

数据内容容错管理
数据历史记录
操作行为记录
定时异地备份

信源同步

信源下发

信源管理数据库

后台抽取

信源验证

数据访问权限管理
远程访问限制
分级角色权限
数据加密存储

信源传输安全管理
Https加密传输
来源IP、访问口令验证





天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发





统一信源管理——全生命周期

配置完成
人工标注
新信源
申请
审核
合格
待配置
不合格
验证
合格
有效信源
不合格
有效
配置有效
无效信源
人工核查
问题信源
配置失效
失效验证
失效
目标失效
失效信源





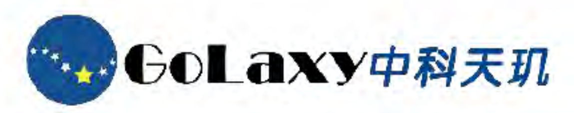
天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
数据资产安全管理	任务调度
信源生命周期管理	智能采集
交互式可视化配置	数据分发





核心技术



统一信源管理——可视化配置

操作简单：

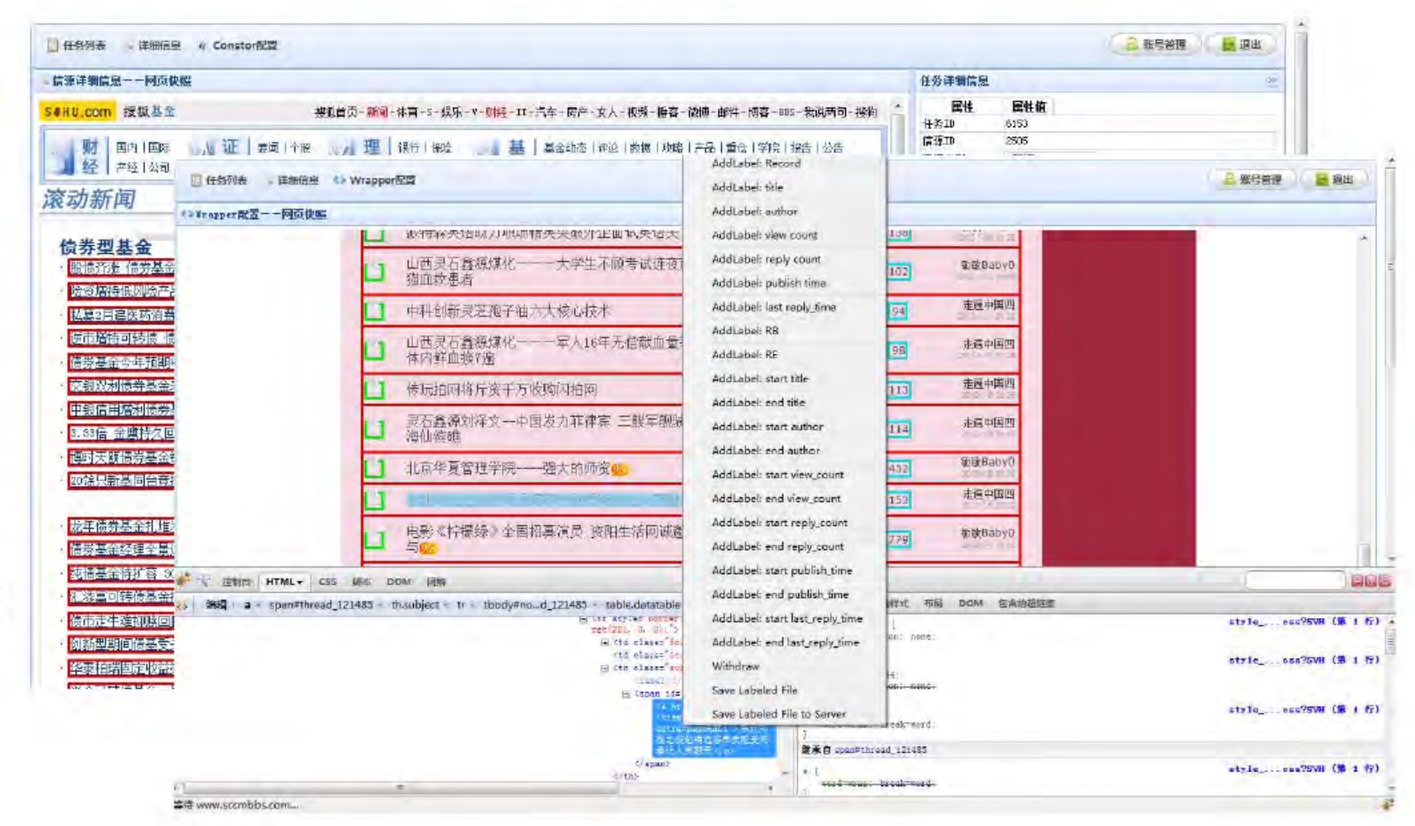
可视化：浏览器可视化界面

交互式：鼠标点击即可完成

安装简单：

B/S结构：随时随地在线完成

插件依赖：拖拽式一键安装





核心技术



天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
傻瓜式可视化配置	任务调度
信源生命周期管理	智能采集
数据资产安全管理	数据分发







统一信源管理——整体架构

基于Browser的配置前台

基于Java的Web Server

数据采集

信源数据

信源同步

信源下发

信源管理数据库

后台抽取

信源验证





分布式数据采集

数据服务接口

数据过滤分发

IP 池服务

隧道服务

JS解析服务

截图取证服务

智能采集

任务调度

任务队列

资源调度

信源数据

信源同步

调度配置

通道信源配置





天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
傻瓜式可视化配置	任务调度
信源生命周期管理	智能采集
数据资产安全管理	数据分发



核心技术



分布式数据采集——资源调度

资源调度：

管理、调度、监测可用运算资源；
以采集器为单位，对采集资源进行调度、监测；
实现负载均衡、高可用性

支持：

1. 持续任务
2. 定时任务
3. Cron任务

Registry(zk, redis)
register subscribe
JobClient
node
master
NodeGroup...
FailStore
submit
feedback
retry
JobTracker
node
master
NodeGroup
QueueManagers
pull
push
finish
resend
TaskTracker
node
master
NodeGroup...
FailStore
management
JobLogger
CronJobQueue
ExecutableJobQueue
ExecutingJobQueue
FeedbackJobQueue
management
Admin (Web)





分布式数据采集——资源调度

开始

JobClient
生成任务

JobTracker
分发任务

TaskTracker
执行任务

任务
队列

Registry(zk, redis)

register
subscribe
register
subscribe
register
subscribe

JobClient
node
master
NodeGroup...
FailStore

JobTracker
node
master
NodeGroup
QueueManagers

TaskTracker
node
master
NodeGroup...
FailStore

submit
feedback
retry
pull
push
finish
resend
management

JobLogger
CronJobQueue
ExecutableJobQueue
ExecutingJobQueue
FeedbackJobQueue

management

Admin (Web)



核心技术



天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
傻瓜式可视化配置	任务调度
信源生命周期管理	智能采集
数据资产安全管理	数据分发







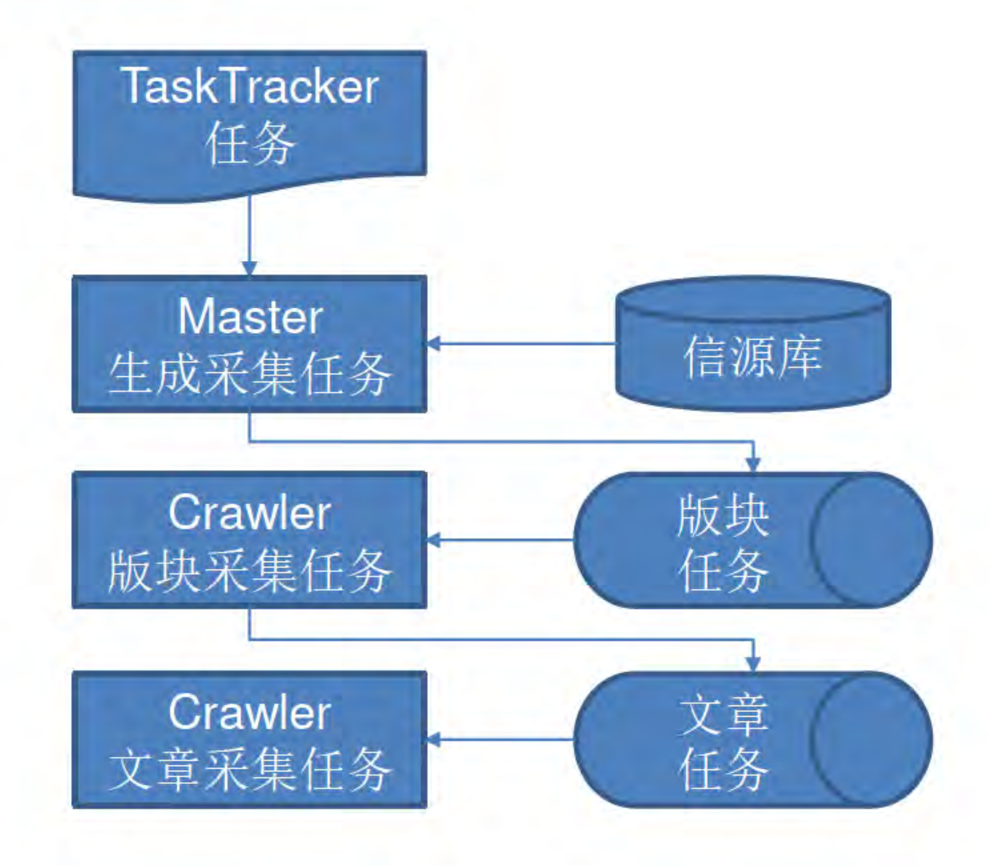
分布式数据采集——任务调度

任务调度：

管理、调度、监测采集任务；

以信源为单位，对采集资源进行调度、监测；

实现分级优先采集





核心技术



天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
傻瓜式可视化配置	任务调度
信源生命周期管理	智能采集
数据资产安全管理	数据分发

核心技术



分布式数据采集——智能采集

智能采集：

按需执行、监测采集任务；

智能防封堵、隧道穿透、动态解析；

实现高召回、高准确数据采集抽取

TaskTracker 任务

IP 服务

隧道服务

Js加载服务

日志分析服务

Master 生成采集任务

Crawler 版块采集任务

Crawler 文章采集任务

信源库

版块任务

文章任务

日志信息

调度信息

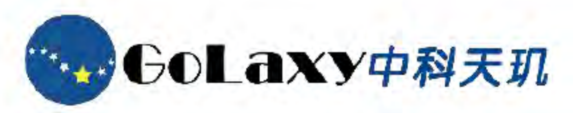




分布式数据采集——智能采集

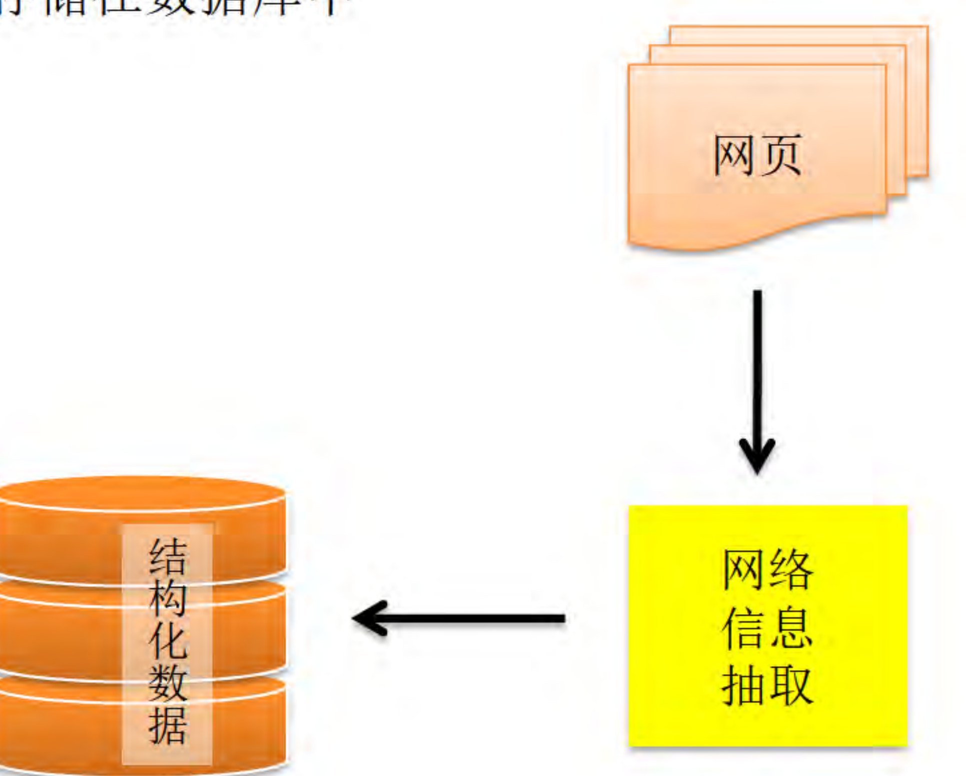
覆盖	支持	采集	保留
境内外网络新闻	网页	正文	原始快照
新闻APP	文档	评论	
微信	多媒体	跟帖	
境内外微博		点击数	
贴吧论坛博客		评论数	

核心技术



分布式数据采集——智能采集

- **信息抽取：** 从半结构化的网页中抽取出数据，结构化地存储在数据库中
- **技术方案**
  - 半自动化方法
    - 手工编写模板
  - 自动化方法
    - 基于统计的方法
      - 构建DOM树，利用链接文字比、文字密度等统计信息
- **核心组件**
  - 通用信息抽取工具Mask
  - 搜索引擎结果抽取组件ARISE
  - 多记录网页内容抽取组件Constor
  - 单记录网页内容抽取组件HtmlSAE
  - 网页的同模板自动检测组件WrapperCheck



核心技术



分布式数据采集——智能采集

抽取组件主要用于资讯类新闻网页抽取，标准的页面结构抽取准确。

抽取组件提供其他功能

抽取组件提供编码识别，编码转换。语种识别，

抽取组件支持抽取多语言

抽取组件支持多语言抽取（英语，中文，藏语，日语，韩语，德语，阿拉伯语）

抽取组件准确率

抽取组件准确率包括正文、标题、发布时间等准确率。综合准确率可以达到85%。



核心技术



天玑数据采集中心

统一信源管理	分布式数据采集
多租户集中管理	资源调度
傻瓜式可视化配置	任务调度
信源生命周期管理	智能采集
数据资产安全管理	数据分发







分布式数据采集——数据分发

产品及服务



天玑数据采集中心建有数据实时推送系统、历史数据回溯系统满足用户对于实时数据和历史数据采集需求，同时也可提供元搜索工具或轻便的定制化采集小工具来满足客户定制化的采集需求。









专题数据推送

基于关键词规则和信源筛选的按需定制和推送

全量数据推送

全量基础数据实时推送

数据回溯工具

7种工具回溯微博、微信、贴吧、新闻的账号、文章数据

元搜索工具

支持百度、必应、搜狗、360、谷歌、微信、贴吧、微博等常规搜索引擎和社媒网站搜索采集

采集小工具

轻便、部署简单、扩展方便、功能齐全

调度中心、信源配置中心、采集执行器三大组件，均支持分布式，支持HA

灵活的采集模板配置，解决大部分采集场景需求

支持多线程采集、动态代理、失败重试等



实时数据推送系统



主题数据采集

通过关键词及媒体的筛选方式组合，过滤出符合配置条件的特定数据。

配置方式：

关键词配置
媒体配置
关键词+媒体配置

数据实时推送系统　当前用户： quanliang　退出

消费状态
令牌管理
帮助
数据字典
REST接口文档
WEBSOCKET接口文档

当前时间：2018年8月27日18:40:00

消费状态

新闻　论坛　博客　微博　微信　境外新闻　APP　电子报　问答　贴吧　Facebook的　推特

数据产生记录

起始位置：12114447　结束位置：12281806

当前位置：12114447

数据产生记录

日期	实际产生数据	日最大可产生

文件请求

2018-08-26 ～ 2018-08-27　查询

取消　保存



数据回溯工具



对实时数据采集范围之外的历史数据的采集，包括微博关键词采集、微博账号回溯、微信公众号回溯、微信文章监测、新闻关键词采集、贴吧关键词采集共7种采集方式

数据服务系统

首页 热点榜单 数据实时推送系统 分析报告 数据回溯工具 登录 试用申请

- 微博关键词采集
- 微博账号回溯
- 微信关键词采集
- 微信公众号回溯
- 微信文章监测
- 新闻关键词采集
- 贴吧关键词采集

输入项示例

关键词规则：五月散人

关键词规则：多个关键词以空格分隔，如：“台湾 大选”，关键词间关系为与。关键词只可由汉字、英文、数字组成，每个关键词字符都不得少于2。

采集时段：2018-06-04 ~ 2018-06-04

开始采集

采集结果示例

采集数据结果：

可获得用户名称、文章内容、发布日期、发布时间、转发数、评论数、点赞数、命中关键词、倾向性、分词、原始链接，和是否原创字段。

元搜索工具



搜索引擎采集软件

搜索引擎采集软件

- 搜索
- 任务
- 微信公众号
  - 微信搜索
  - 微信公众号
- 统计监控
  - 统计
  - 监控
- 设置

微信公众号　　公众号名称/账号　查询　新增公众号

其实健康很简单　　上次更新时间：2020-01-02 09:29:20
微信号：40jiandan8322
微信介绍：jiandan8322免费口腔黏膜病健康咨询，查口腔黏膜疾病，找口腔黏膜专业医生，了解口腔黏膜病全面知识，热线:01082062598　删除

会说话是一种教养　　上次更新时间：2020-01-02 09:29:33
微信号：40gh_619b3fa8143c
微信介绍：你那么优秀，别输在说话上，在这里跟我一起，学做高情商会说话的女人。（商务合作请加QQ：3461483413）　删除

甘南鼎兴传媒　　上次更新时间：2020-01-02 09:29:58
微信号：40znzdcr931021
微信介绍：znzdcr931021甘南鼎兴传媒公众平台是发布便民服务、生活信息、招聘求职信息、企业宣传为一体的综合服务平台。　删除

叫我小帆帆　　上次更新时间：2020-01-02 09:30:11
微信号：40gh_93d91417de22
微信介绍：关注我，每天跟你分享你喜欢的故事　删除



定制化采集小工具



golaxy-crawler 是一个分布式定制化采集工具，其核心设计目标是使用简单、 功能齐全、模板灵活、轻量级、易扩展， 开箱即用。

信源管理：网站配置、模板配置、处理器配置、项目管理、环境管理、用户管理

调度管理：运行报表、任务管理、调度日志、采集执行器管理

运行配置：动态参数、动态代理、失败重试、主动停顿、全局去重、……

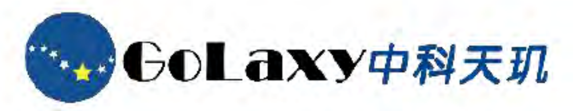
采集节点：Internet、数据处理器、页面解析器、页面加载器、……、采集线程池、待采集任务池、已采集任务池、运行时数据

……

采集节点：Internet、数据处理器、页面解析器、页面加载器、……、采集线程池、待采集任务池、已采集任务池、运行时数据

数据输出：Kafka、Elasticsearch、Mongodb、文件、控制台

定制化采集工具模板

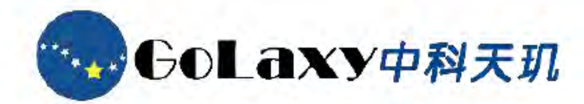


采集模板共包含entries、attributes、mapping、strategy、storage、handler六大部分，控制自定义采集流程：

- Entries：采集任务入口列表，必须，格式为json数组，包含urlPattern、翻页参数pages、http请求参数、自定义属性attributes；
- Attributes：任务携带属性，可选，json格式，可以将entries中所有任务的共性属性提取出来放到这来配置；
- Mapping：字段信息提取映射关系，必须，json格式，包含提取方法method、行数rows、字段列表fields；
- Strategy：主键和去重策略，可选，主键（_id）根据key、algorithm、prefix三个字段生成；
- Handler：数据处理器，可选，在数据最终写入数据源之前对数据做进一步处理；
- Storage：存储设置，可选，指定存储数据源类型和目标库表

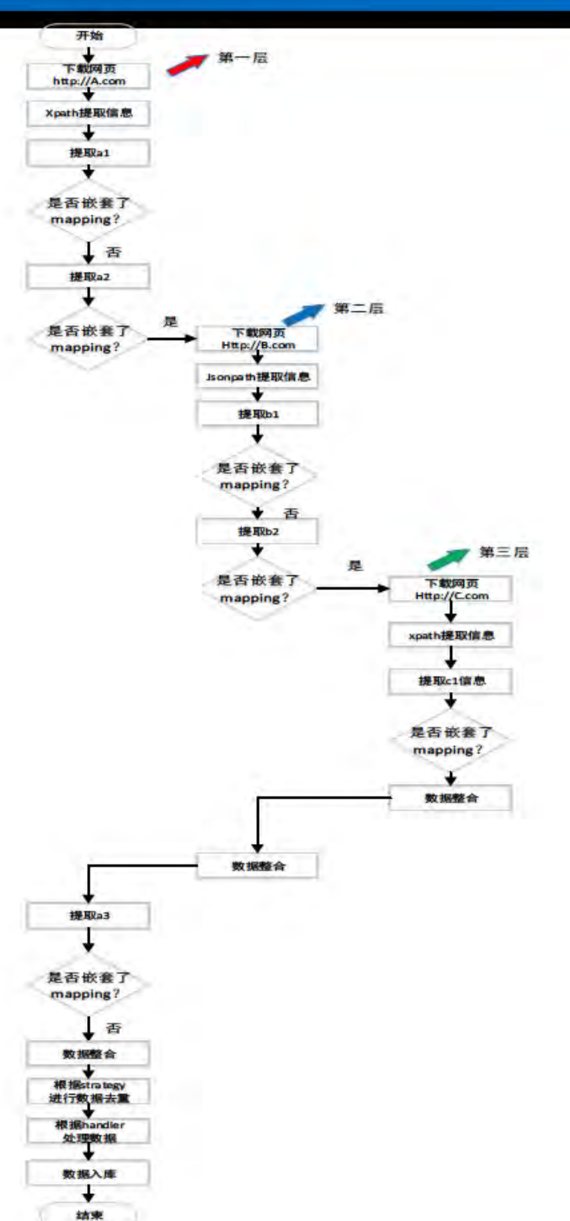
采集模板
entries
urlPattern
pages
httpParams
attributes
attributes
strategy
key
algorithm
duplicates
expired
prefix
storage
ds
handler
name
type
mapping
method
rows
fields
key
type
format
value
required
multivalue
store
mapping
mappingEntry
attachmentEntry

定制化采集工具模板

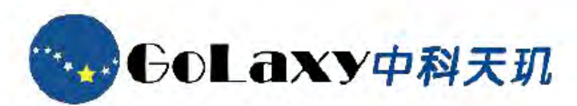


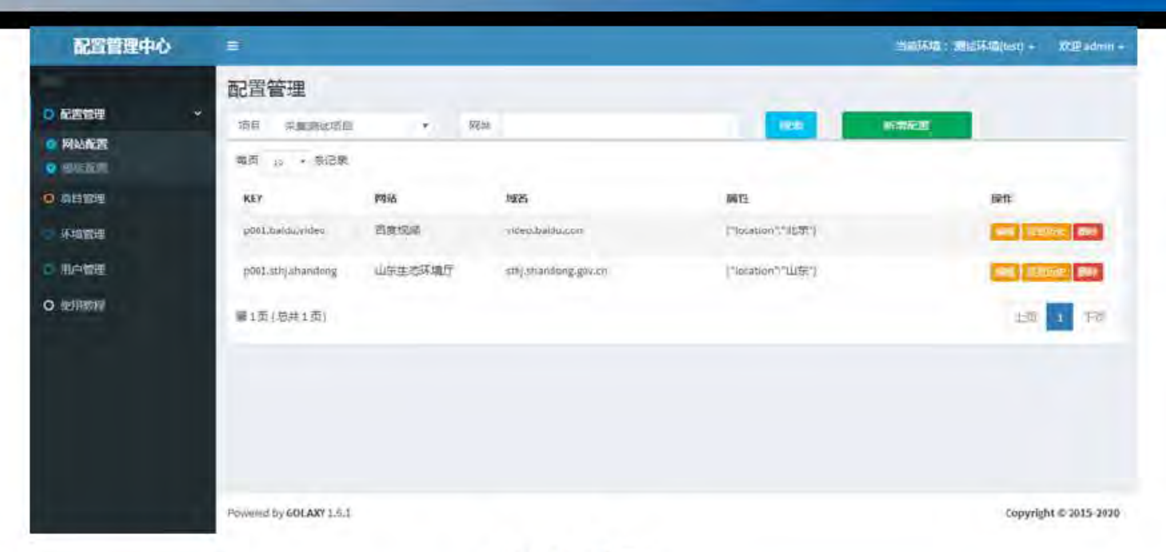
采集模板配置灵活，可通过采集模板控制程序采集流程，如下：



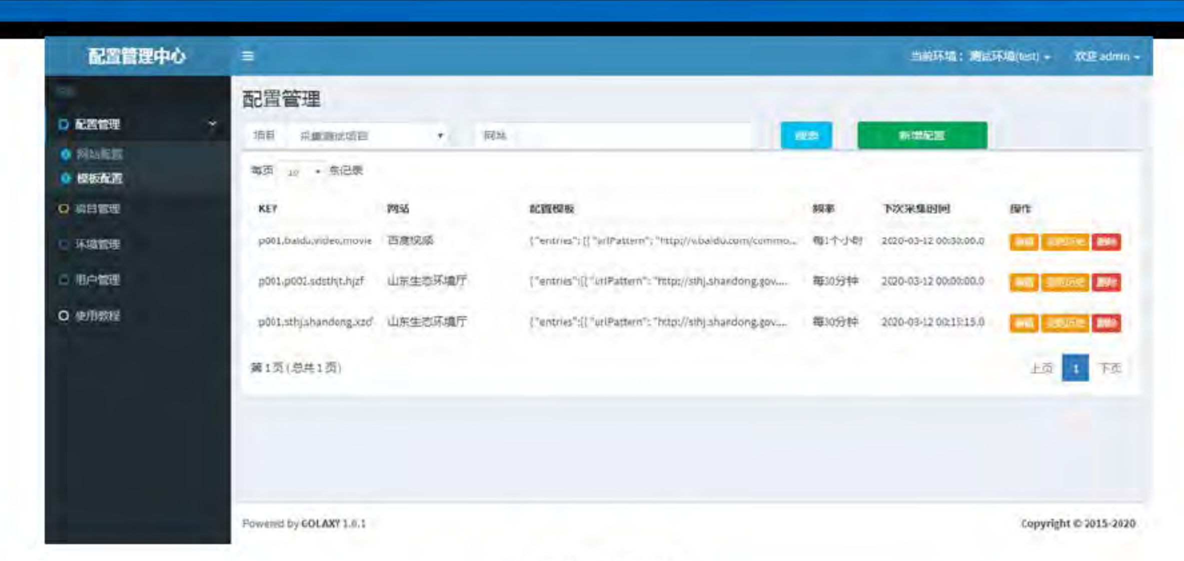


定制化采集工具-配置展示

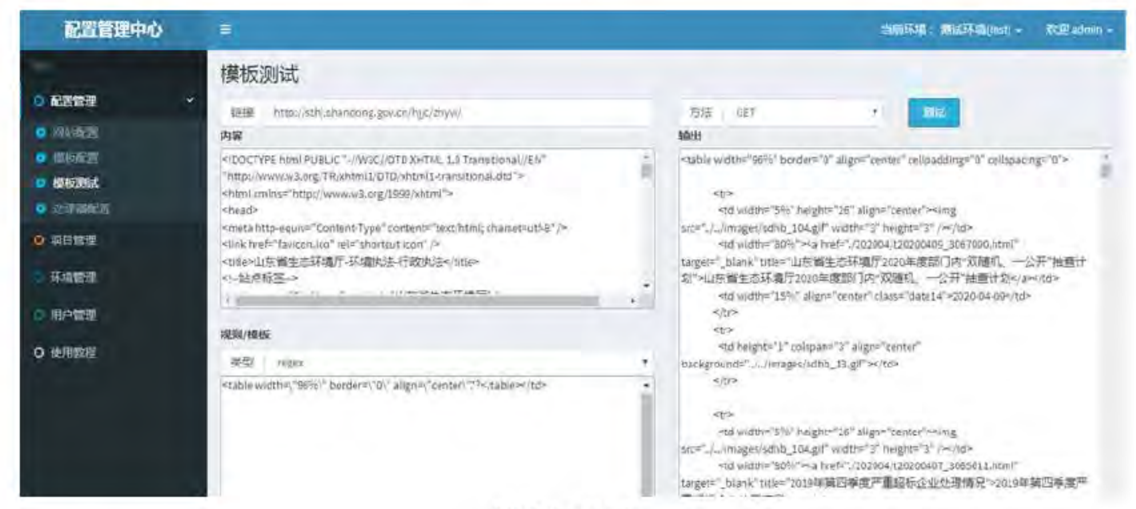




网站配置



模板配置



模板测试



处理器管理



感谢您的观看

中科天玑数据科技股份有限公司

序号	成果名称	概述	技术状态（在研、测试、试验、验收、列装等）	主要用户
1	社交机器人检测	面向国内外主要社交媒体平台，基于特征工程、图神经网络、群组分析等技术，构建社交机器人样本库、话术库、群组关联关系等知识库，识别涉我敏感话题机器人	列装	中央网信办
2	智能算法缺陷分析	针对推荐算法、调度决策算法、检索过滤算法、精准排序算法、生成合成算法这五类常见信息服务算法，研究算法机制、机理，对算法的多样性、公平性、可信性等方面进行风险监测和安全评估	列装	中央网信办
3	大模型攻防技术	研究大模型的数据接入、风险监测、模型脆弱性分析、测试数据集自动构建、可信性测评等	在研	中央网信办
4	智能宣传辅助系统	采集传统媒体、社交网站、即时通信等800万个涉及政治、军事、外交、经济数据信源数据，进行事件分析、人物/群体画像，通过自有账号群建设，内容生成，跨平台智能推送至Twitter、Facebook等社交媒体实现宣传辅助	验收	军委科技委
5	代理实验平台系统	为使用各类境外社交媒体平台的用户或者机构提供批量账号培育服务，提供社交媒体平台账号代理，实验场景编排，实验场景执行等能力	验收	军委科技委
6	智能投送系统	面向批量化的社交媒体账号，提供具备精确传播致效能力的大规模调度服务，目前已实现资源调度、行动规划、协同调度等模块，支持万级万级账号并发调用，千级任务并发	在研	311
7	选情分析系统	面向M、T、港等主要地区，有效支撑对M2016年、2020年，T2020年地区领导人选举，香港2022年特首选举，相关选情系统及选举报告获重要批示，目前支持T2024年地区领导人选举工作	在研	国台办

编号:

密级

2020年11月23日 星期一

批量调度策略限制

发布，评论，转发

- 消息和Bot关系
  - 一条消息可以被多个Bot同时发布
  - 一条消息不可以在4个小时内被一个bot重复发布
  - 一个bot发布一条消息时，bot在5分钟内不能执行其他任务
- bot占用：
  - 任务对bot的占用为五分钟
  - 在业务层需要将间隔扩展为十分钟

点赞，关注：

每个bot针对同一个用户或消息只能执行一次

点赞和关注操作由于IP过少，账号特性驱动，时间间隔需要在小时级别

智能引导组2021年度规划

王永庆

智能引导组定位

- 承担核心系统研制、关键技术攻关以及重要纵向课题申报
  - 1套核心系统：GoPro（智能引导系统）
  - 4组关键技术：智能创作、观点挖掘、账号培养、用户画像
  - 横纵向课题申请：网安、网信、KJW等来源方向
  - 纵向课题支撑：网安、网信、KJW等来源方向

系统目标

- GoPro系统：
  - **核心定位：**以“人”为核心的宣传操作平台
  - **核心价值：**智能辅助境外定向宣传
  - **系统特点：**交互式观点内容生成、事件倾向性观点分析、高价值账号培养、跨平台用户分析

发展规划

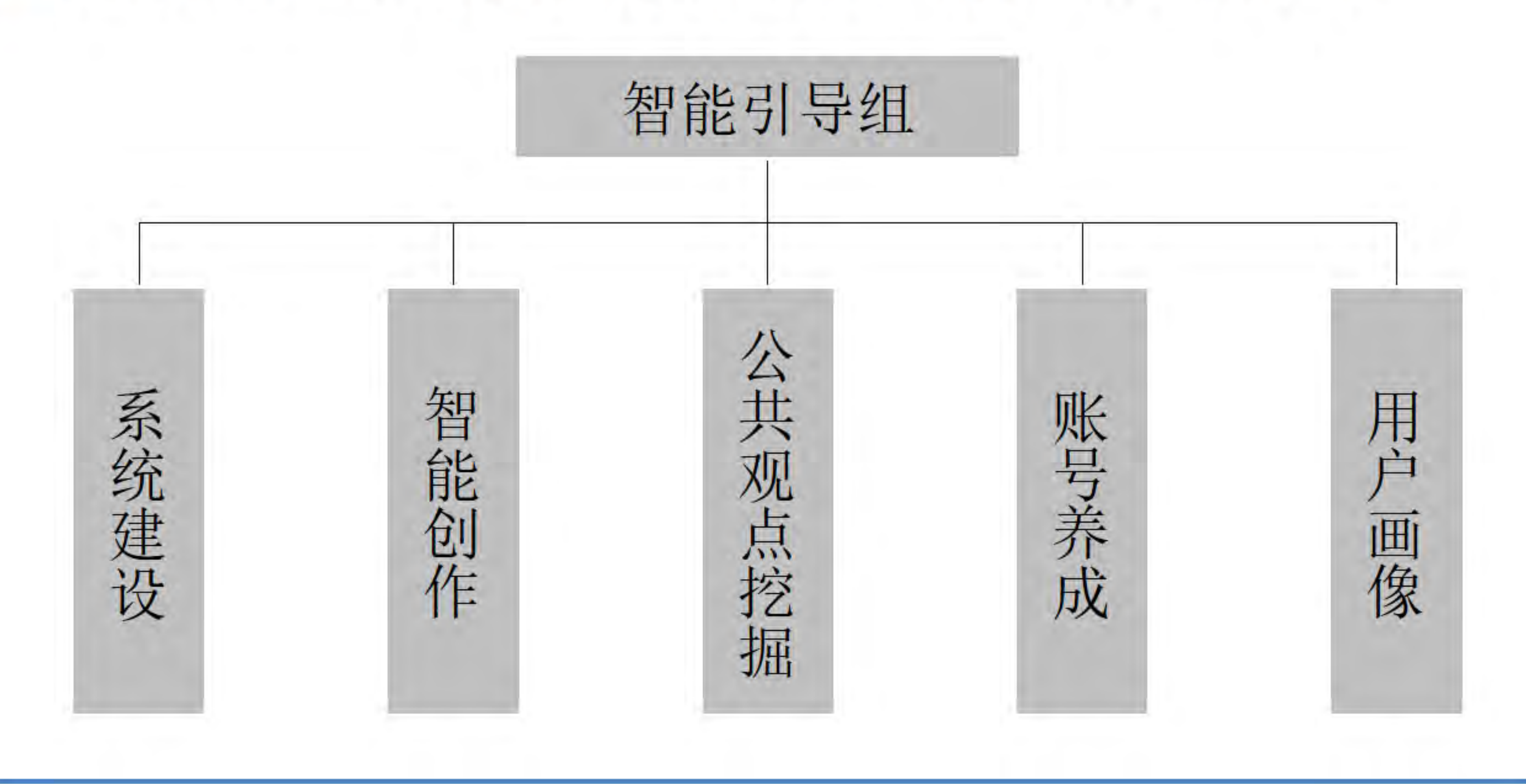
- 第一阶段（2020-2021）：内部发展，确定智能引导系统的基本系统形态及框架
- 第二阶段（2022-2023）：外部生存，确定智能引导系统的核心用户群，关键技术就具体应用升级
- 第三阶段（2024-）：影响力提升，立足国家级项目，申报国家级奖项

年度成果预期

- 专利/软著10项：涉及数据服务、业务通信、权限管理、文本生成、观点挖掘、账号培养、用户画像等方面
- 高水平论文10篇：涉及文本生成、观点挖掘用户画像等方面
- 开源算法平台1套
- 培育方向4个

智能引导组设置

- 围绕GoPro系统能力建设，预设5个方向：



各方向人员安排

- GoPro系统建设方向：
  - 方向负责人：王清林
  - 成员：刘金虎、朱川生、薛雪
- 智能创作与推送策略优化方向：
  - 方向负责人：庞亮
  - 成员：丁汉星、侯丹阳、朱运昌、郝长盈、肖飞

各方向人员安排

- 公共观点挖掘与事件分析方向：
  - 方向负责人：高金华
  - 成员：赵睿卓、王亚杰、姜燕、柯文俊、赵鑫

- 用户画像与群体分析方向：
  - 方向负责人：王永庆
  - 成员：段雪野、高浩、邵江立、张皓彭、史铂深

各方向人员安排

- 账号养成与交互策略优化方向（培育）：
  - 方向负责人：孟剑

新风貌 新期望

- 正一点：风气正
- 硬一点：能力硬
- 飒一点：朝气蓬勃

愿未来可期！

用户画像系统目标围绕重点地区收集互联网数据绘制具有一定地区覆盖度的社交用户画像。现阶段主要围绕香港地区的 Twitter 用户，目标绘制 20 万（Twitter 的香港用户约 75 万）用户的社交用户画像。画像具体包括用户的 Twitter 账号信息及用户性别、年龄、职业、所在地、教育水平的基础用户属性。

当前系统通过与中科天玑、郑州分所、采集组合作，以及通过组内自采集方式累积微博用户账号 45,302,104 条，内容信息 434,230,549 条；Twitter 用户账号 32,148,771 条，内容信息 245,799,086 条。按照用户数据类型，将采集数据按照用户属性、行为、内容、关系进行分类处理，共累积 Twitter 用户属性 3200 万+、行为 2.4 亿+、内容 2.4 亿+、关系 950 万+；微博用户属性 4500 万+、行为 1 亿+、内容 4.3 亿+、关系 220 万+。

按照四类用户数据设计数据服务访问接口共 36 个，以上全部数据可通过接口访问。通过数据访问服务接口，设计实现用户画像检索系统 PYou Twitter 及微博版，内网访问地址：10.61.1.241:5000，提供通过用户昵称、所在地、内容关键词及 hashtag 检索用户数据。在已有积累数据基础上，对用户账号的归属地进行推断，归拢所有可能的香港地区用户账号，目前已累积可能的香港地区 Twitter 用户账号 17 万+，具体的属地推断结果展示可访问内网地址：10.61.150.2:5005。

此外，目前还初步探索用户画像的具体展现形式，设计实现用户属性、行为、内容、关系四类数据的展现，具体可访问：10.61.150.2:5005，点击用户详情信息即可查看。

系统截图如下：

- PYou 用户画像系统：10.61.1.241:5000





Twitter用户昵称　Apple

查询用户：Apple

20 条/页　首页　上一页　下一页　末页　当前第 1/150 页，共2985 条记录

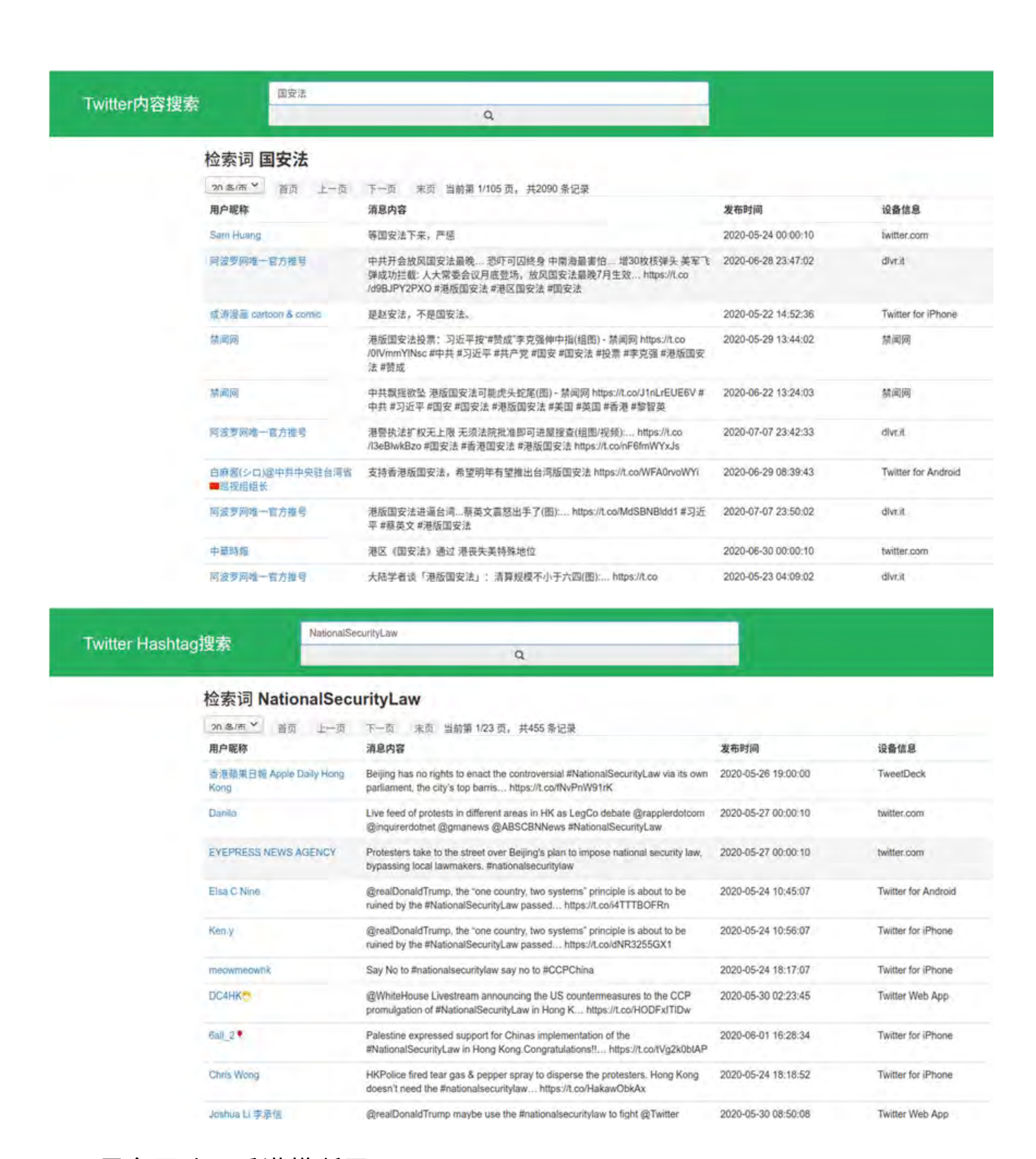
头像	昵称	位置	性别	粉丝数	关注数	操作
	Apple Apple		未知	16465	45	查看用户主页
	用户描述	中部精油舒壓芳療 單純按摩無約炮 沒有外跑、沒有到府按摩只能你們過來我這 最晚前一天預約不接凌晨按摩一律先匯款。				
	Apple Apple		未知	0	32	查看用户主页
	用户描述					
	Apple Apple		未知	-1	-1	查看用户主页
	用户描述					
	APPLE APPLE APPLE	SinRin Nation	未知	-1	-1	查看用户主页
	用户描述					
	Apple		未知	-1	-1	查看用户主页
	用户描述					

Twitter用户所在地　香港

查询地点：香港

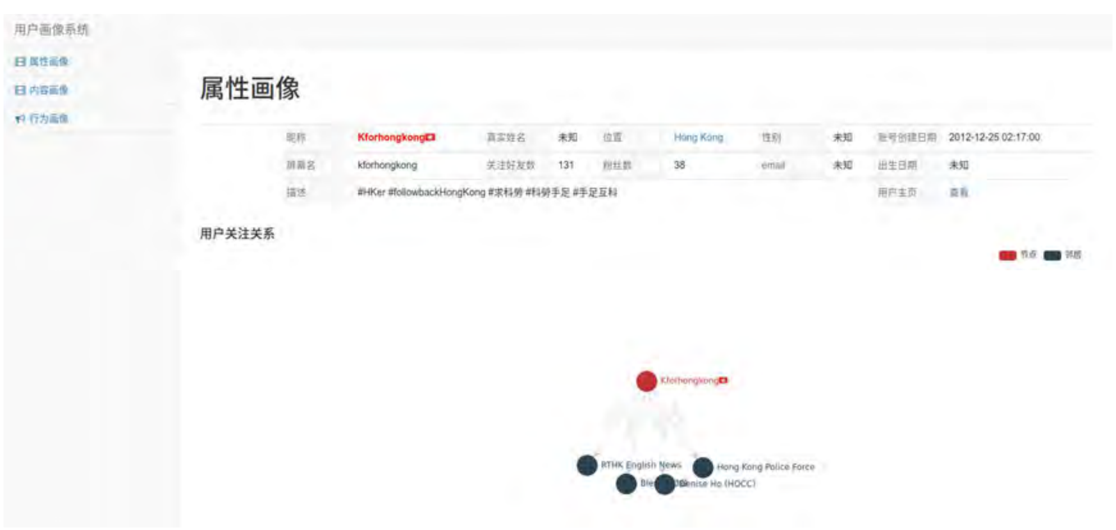
20 条/页　首页　上一页　下一页　末页　当前第 1/ 500 页，共13680 条记录

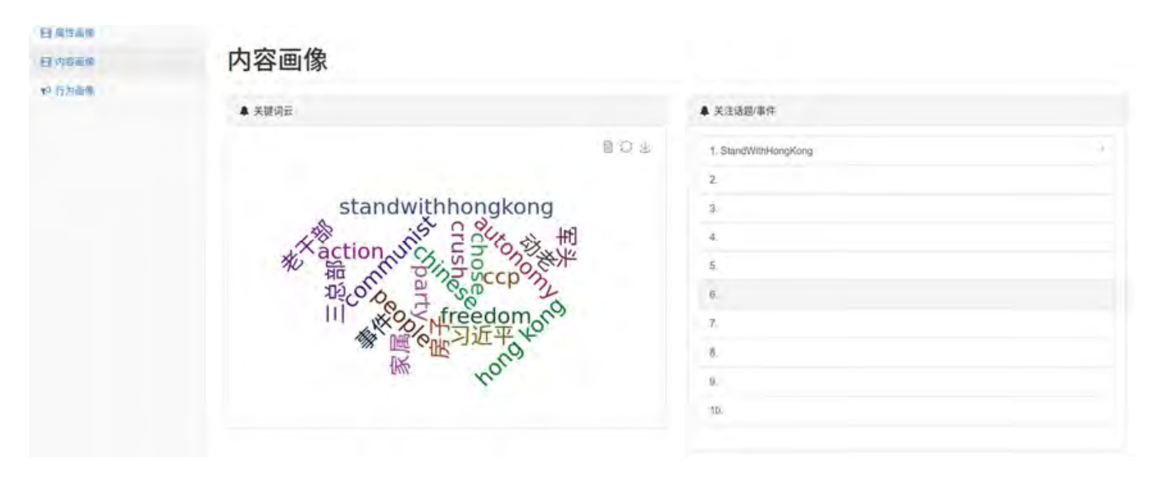
头像	昵称	位置	性别	粉丝数	关注数	操作
	但丁	香港香港	未知	-1	-1	查看用户主页
	用户描述					
	Hui Man-cheong	香港	未知	-1	-1	查看用户主页
	用户描述	http://Godahsing.com總編輯．股東倡議活動者．新創公司菜鳥．專業投資者．節目主持．發言人．作者．男				
	伊死	香港	未知	-1	-1	查看用户主页
	用户描述	元編集長に名乗った元雑役。現、某車ケアショップの通訳。声優厨、漫画厨。珈琲、マンガと声優がないと死ぬ。MONSTER HUNTERに永遠のHR2ハンター。TweetはPlurkから転送。				
	iwkkwi	香港	未知	13	16	查看用户主页
	用户描述	Action figure collector Toy photographer sketch				
	吾奕信	香港	未知	-1	-1	查看用户主页
	用户描述	一個生於異國他鄉的東西南北人。				



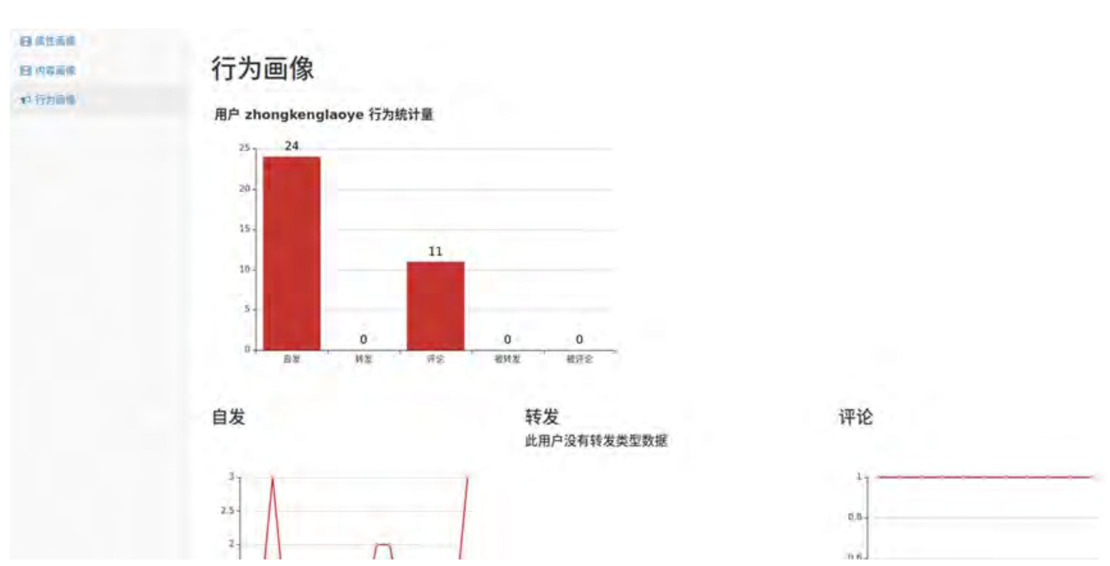
● 用户属地—香港推断展示：10.61.150.2:5005


用户画像
20 条/页
首页
上一页
下一页
末页
当前第 1/8773 页，共175449 条记录
昵称
用户名
性别
香港人概率
用户描述
操作
GA_hker
hongkonger
未知
1.0000
Hker,Free Hong Kong ,Anti extradition law, Anti Chinazi twitter綠要打 YouTube 綠都要打 #StandwithHK #hongkongers #FreeHK #手足互科 || https://hkrev.info/main_en.php ||
查看用户主页
用户填写地区：Hong Kong
地区匹配：是，地区匹配得分：0.8250
用户名匹配：是，用户名匹配得分：0.9325
用户描述匹配：是，描述匹配得分：0.9950
用户语言习惯匹配：否，语言使用习惯匹配得分：0.0000
AmyKong75018499
Kong
未知
1.0000
香港人HKer 光復香港
查看用户主页
用户填写地区：Hong Kong
地区匹配：是，地区匹配得分：0.8250
用户名匹配：是，用户名匹配得分：0.5750
用户描述匹配：是，描述匹配得分：0.9994
用户语言习惯匹配：否，语言使用习惯匹配得分：0.0000
hker_coco
Cocochc
未知
1.0000
HKer / Stand with Hong Kong
查看用户主页
用户填写地区：Hong Kong
地区匹配：是，地区匹配得分：0.8250


用户画像系统
属性画像
内容画像
行为画像
属性画像
昵称
Kforhongkong
真实姓名
未知
位置
Hong Kong
性别
未知
账号创建日期
2012-12-25 02:17:00
屏幕名
kforhongkong
关注好友数
131
粉丝数
38
email
未知
出生日期
未知
描述
#HKer #followbackHongKong #求科勞 #科勞手足 #手足互科
用户主页
查看
用户关注关系
Kforhongkong
RTHK English News
Hong Kong Police Force
Denise Ho (HOCC)


属性画像
内容画像
行为画像
内容画像
关键词云
standwithhongkong
老干部
action
communist
chinese
crushe
chose
autonomy
动老头
羊头
ccp
三总部
party
people
事件
家属
房子
freedom
习近平
hong kong
关注话题/事件
1. StandWithHongKong
2.
3.
4.
5.
6.
7.
8.
9.
10.


情感走势
活跃度曲线
负面
中性
正面
50
40
30
20
10
0
08-04 2020
08-06 2020
08-08 2020
08-10 2020
08-12 2020
08-14 2020
08-16 2020
08-18 2020
08-20 2020
语种分布
语种分布
英语
塔加路语
中文(简体)
最新微博
Henry
RT @SecPompeo: The Chinese Communist Party chose to crush the freedoms and autonomy of the people of Hong Kong. Because of the CCP's action...
2020-06-21 06:51:58
Henry
@iingwen #StandWithHongKong #香港加油
2020-07-01 01:07:32
Henry
RT @HowChinaToday: 大事件！习近平开始动老军头们的房子了。三总部的老干部和家属们都炸锅了！ https://t.co/j1nJ38pdpZ
2020-06-28 19:32:32


属性画像
内容画像
行为画像
行为画像
用户 zhongkenglaoye 行为统计量
24
0
11
0
0
自发
转发
评论
被转发
被评论
自发
转发
此用户没有转发类型数据
评论

用户相关性排行榜

uid	all	用户转发	用户评论	转发用户	评论用户
Onebtcer	1	0	1	0	0
MsVChina1	1	0	1	0	0
galileocheng	1	0	1	0	0
Daisy_Zhou	1	0	1	0	0
waylybaye	1	0	1	0	0

用户影响力最大内容排行榜

内容id	影响力因子

